*Bioinformatics*, 34(24), 2018, 4284–4286
doi: 10.1093/bioinformatics/bty510
Advance Access Publication Date: 22 June 2018
Applications Note





Genome analysis

# MetaPOAP: presence or absence of metabolic pathways in metagenome-assembled genomes

**Lewis M. Ward[1,*], Patrick M. Shih[2,3,4] and Woodward W. Fischer[5]**

[1]Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA 02138, USA, [2]Environmental Genomics and Systems Biology Division, E O Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [3]Feedstocks Division, Joint Bioenergy Institute, Emeryville, CA 94608, [4]Department of Plant Biology, University of California, Davis, CA 95616, USA and [5]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 19, 2018; revised on June 18, 2018; editorial decision on June 19, 2018; accepted on June 20, 2018

## Abstract

**Summary:** Genome-resolved metagenomics allows the construction of draft microbial genomes from short-read shotgun metagenomics (Metagenome-Assembled Genomes, or MAGs); however, even high-quality MAGs are typically somewhat incomplete and contain a small amount of contaminant sequence, making accurate prediction of metabolic potential challenging. Here, we describe MetaPOAP, an algorithm for probabalistic assessment of the statistical likelihoods for the presence or absence of metabolic pathways in MAGs.

**Availability and implementation:** MetaPOAP is available as Python scripts on GitHub or from the Fischer lab webpage, https://github.com/lmward/MetaPOAP.

**Contact:** lmward@post.harvard.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

With the explosion of genome-resolved metagenomic sequencing, most available microbial genomes are MAGs—Metagenome-Assembled Genomes—produced through assembly of high-throughput, short-read metagenomic data into contigs and binning of those contigs into draft genomes (Bowers *et al.*, 2017). MAGs have become a valuable resource for investigating the diversity, metabolic potential and evolutionary history of environmental microbes (e.g. Anantharaman *et al.*, 2018; Parks *et al.*, 2017; Ward *et al.*, 2018); however, MAGs have varying degrees of completeness, and often include some amount of contaminant sequence (e.g. Bowers *et al.*, 2017; Parks *et al.*, 2017). For many environmental, ecological and evolutionary studies it is desirable to draw inferences about the absences of particular genes or metabolic pathways from a given MAG. However, due to processes required to construct MAGs from short-read sequence data it can be unclear whether the absence of genes is an authentic signal of their absence from the source genome, or if instead these genes were present in the source genome but not recovered in the MAG (i.e. genes that reside in an incomplete portion of the genome), giving a False Negative result. Here, we describe an application for estimating the likelihood that a metabolic pathway is truly absent from a genome given a MAG of known degree of completeness; this approach provides a measure of False Negatives. In addition, automated genome binning procedures typically recruit some number of contigs that do not belong to the source genome (i.e. contamination); sometimes this can be mitigated by careful downstream curation of genome bins. To aid in this, we also designed a means for estimating the likelihood that a gene or metabolic pathway was mistakenly recruited to the MAG (i.e. belongs to the contaminant fraction); this constitutes a metric for False Positives. These tests are incorporated into a program called MetaPOAP (pronounced like soap, mɛtə-poʊp), the Metagenomic test for Presence Or Absence of Pathways. This program uses inputs of MAG completeness and contamination, number of coding sequences, the total number of key marker genes in the complete pathway of interest, and the number of genes in the pathway recovered in the MAG as well as the number of contigs on which they were found, and makes a probabilistic calculation of the likelihood

of False Negative and/or False Positive results, respectively, for a specific metabolic pathway of interest. Adding statistical insight into the true absence of genes and pathways fills an important niche until sequencing technologies and assembly processes improve to the point that complete MAGs are the norm.

## 2 Implementation

MetaPOAP produces estimates of probability of missing genes from a pathway (False Negative estimate) and the probability of genes in a pathway having been mistakenly assigned to a MAG (False Positive estimate). These calculations are performed by Python scripts, following mathematics first described in Ward *et al.* (2018) and summarized below.

The complete workflow for MetaPOAP estimates is described in detail in the Supplementary Material. In brief, prior to estimation of False Negative and False Positive statistics by MetaPOAP, MAGs must be analyzed to estimate completeness and contamination. In particular, MetaPOAP is currently implemented to utilize these estimates as output by CheckM (Parks *et al.*, 2015). The number of recovered coding sequences must also be calculated, such as by Prodigal (Hyatt *et al.*, 2010) or as part of an annotation pipeline such as RAST (Aziz *et al.*, 2008). Key marker genes for pathways of interest must be identified and annotated, and the number of recovered contigs encoding these genes must be counted. Care must be taken in assigning the number of marker genes, as paralogous genes can be misannotated as part of a pathway despite physiological use for a different process. For example, most steps in the tricarboxylic acid (TCA) cycle for heterotrophy and the reductive TCA cycle used by some autotrophs are homologous, with either limited (ATP-citrate lyase) or no unique marker genes (reverse citrate synthase) associated with the net carboxylating pathway (e.g. Mall *et al.*, 2018; Nunoura *et al.*, 2018).

The False Negative estimate is formally cast as the probability that certain genes occur in a genome of interest but were not recovered in the MAG. To calculate this, MetaPOAP estimates the probability mass function of recovering zero genes of a particular set from a genome of predicted size, given independent estimates of completeness, and assuming random sampling without replacement of individual genes. As sequence length is more meaningful than contig-level colocalization for unrecovered sequence, the False Negative estimate is made on the gene level assuming average gene length. Though gene size varies (e.g. Milo and Phillips, 2015), this approach simply assumes that all genes in the pathway of interest are of approximately common length (i.e. encoding an average protein length of ~270 amino acids, Brocchieri and Karlin, 2005) and have an equal probability of being selected. Under these assumptions, operon structure and colocalization of genes can be disregarded for the False Negative estimate. Given the way that this is calculated, a high False Negative estimate formally means that the likelihood of failing to observe a set of genes is high even if it is in the source genome; it does not require that the genes necessarily are present—only that the hypothesis that they're present can't be falsified.

The False Negative calculation takes the form of $f(x) = \binom{n}{x}\binom{T-n}{r-x}/\binom{T}{r}$, where $f$ is the probability of recovering $x$ genes of set $r$ from a genome made up of $T$ genes of which $n$ were recovered. $T$ is estimated based on the number of genes recovered, the estimated completeness of the MAG, and assuming an average size distribution of gene length. Regardless of how many marker genes for the pathway of interest were recovered, the False Negative estimate calculated by MetaPOAP is the probability that zero of the remaining marker genes were recovered in the MAG assuming their presence in the source genome. In other words, $x = 0$, and $r$ is equal to the total number of marker genes in the pathway of interest minus the number that were confidently recovered in the MAG. In the implementation of MetaPOAP, required input is the number of genes recovered in the MAG ($n$), estimated completeness (used with $n$ to estimate $T$), the number of key marker genes for the complete pathway of interest and the number recovered in the MAG (used together to calculate $r$).

The complementary False Positive estimate (i.e. the probability that all of the contigs encoding genes in the pathway of interest were mistakenly recruited to the MAG and belong to the contaminant fraction) is also encoded in MetaPOAP. Given an estimate of contamination in a MAG, $C$, as assessed by CheckM (Parks *et al.*, 2015) or a similar approach, and the number of contigs recovered with genes in a pathway of interest recovered in the MAG, $k$, the probability, $P$, that all contigs encoding genes in the pathway of interest do not belong to the genome is given by $P = C^k$. MetaPOAP False Positive estimates assume that all contigs in the MAG are equally likely to be contaminants. In some cases, contaminant contigs can be identified via careful analysis of tetranucleotide frequency, GC content, or other characteristics; these are not taken into account by MetaPOAP, and it is recommended that users pursue other applications to refine genomes after MetaPOAP is used as part of an initial screen to determine whether contamination is a concern. Because the contamination estimates provided by programs like CheckM include both contaminant genes placed in a MAG due to strain-level heterogeneity as well as genes from (grossly) unrelated organisms, the resulting False Positive estimate can be taken as a conservative value.

MetaPOAP is provided as three Python scripts, each accepting a different input format. MetaPOAP_i is interactive on the command line to input MAG statistics along with the number of key marker genes in the pathway of interest and the number of these genes that were recovered in the MAG. MetaPOAP_i then outputs the False Negative and False Positive probabilities calculated as described above, along with a text string of whether the pathway of interest is likely present in the source genome. MetaPOAP_cl accepts parameters as command line arguments and provides simple output values in order to more easily be incorporated into data processing pipelines. MetaPOAP_csv takes a CSV file as input (a template is provided with the Supplementary Material) and outputs a CSV file that includes False Negative and False Positive estimates. The source code for all three versions is available and open source, and the underlying calculations can be incorporated into pipelines for MAG analysis to automate this statistical approach as a part of metabolic annotation and inference.

## 3 Discussion

Despite nearly ubiquitous incompleteness and low levels of contamination in published MAGs, statistical inference is rarely employed when using metagenomic data to make interpretations of metabolic potential. MetaPOAP provides a useful utility for ascertaining the probability that a pathway is—or is not—encoded by the genome from which a MAG is recovered. This approach bears on a range of problems, from better predicting the metabolic potential and ecological role of environmental microbes, to more accurately assessing the distribution and evolution of metabolisms across the tree of life. Understanding patterns of presence and absence of metabolic

pathways within a clade can, for instance, be applied to better interpreting the roles of vertical inheritance, loss and horizontal gene transfer in the evolution of phototrophy (e.g. Ward *et al.*, 2018).

It is important to stress, however, that successful use of MetaPOAP relies on thoughtful selection of key marker genes and accurate annotation. As with many automated approaches for working with metagenomic data, misannotation or misidentification of homologous genes can confound conclusions derived from MetaPOAP. There continues to be no substitute for careful oversight of data and analysis.

## Acknowledgements

We thank Daan Speth and Grayson Chadwick for helpful discussion during the development of MetaPOAP.

## Funding

L.M.W. acknowledges support from an Agouron Institute postdoctoral fellowship. W.W.F. acknowledges support of the David and Lucile Packard Foundation and NASA Exobiology award NNX16AJ57G. P.M.S. acknowledge support from a Society in Science–Branco Weiss fellowship from ETH Zurich.

*Conflict of Interest:* none declared.

## References

Anantharaman,K. *et al.* (2018) Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME J.*, **1**. doi: 10.1038/s41396-018-0078-0.

Aziz,R.K. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Bowers,R.M. *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, **35**, 725.

Brocchieri,L. and Karlin,S. (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.*, **33**, 3390–3400.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Mall,A. *et al.* (2018) Reversibility of citrate synthase allows autotrophic growth of a thermophilic bacterium. *Science*, **359**, 563–567.

Milo,R. and Phillips,R. (2015) *Cell Biology by the Numbers*. Garland Science.

Nunoura,T. *et al.* (2018) A primordial and reversible TCA cycle in a facultatively chemolithoautotrophic thermophile. *Science*, **359**, 559–563.

Parks,D.H. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.

Parks,D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533.

Ward,L.M. *et al.* (2018) Evolution of phototrophy in the *Chloroflexi phylum* driven by horizontal gene transfer. *Front. Microbiol.*, **9**, 260.